Gyermeknyelvi korpuszok és erőforrások

Babarczy Anna
egyetemi docens
BME Kognitív Tudományi Tanszék; MTA Nyelvtudományi Intézet
babarczy@cogsci.bme.hu

Tanulmányaimat a University of Edinburgh-ban folytattam elméleti nyelvészetből, kognitív tudományból és mesterséges intelligenciából. Jelenleg a BME Kognitív Tudományi Tanszék tanszékvezető egyetemi docense és a Nyelvtudományi Intézet kutatója vagyok. Kutatási területeim a nyelvfejlődés és a pragmatikai kompetencia kísérletes vizsgálata és számítógépes modellálása.

1. Három gyermeknyelvi korpusz

Jelenleg, azaz 2019-ben, három viszonylag széles spektrumot átfogó, egészében magyar nyelvű – vagy magyar anyagot is tartalmazó – gyermeknyelvi korpuszt ismerünk, melyek: a GABI, a MONYEK és a CHILDES. A GABI (Gyermeknyelvi beszédAdatBázis és Információtár) fejlesztés alatt áll, és (egyelőre) nem elérhető a nagyközönség vagy akár a kutatóközösség számára, de egy rövid ismertető erejéig említést érdemel. A MONYEK (Magyar Óvodai Beszélt Nyelvi Korpusz) regisztrációval kutatási célokra elérhető a MetaShare szolgáltatáson keresztül az alábbi linken: http://metashare.nytud.hu/repository/browse/hungarian-kindergarten-language-corpus/b572a8106ba711e2aa7c68b599c26a06a4db2e695cf94a1cad6bf6793d747d2a/. A CHILDES (Child Language Data Exchange System) magyar és nem magyar nyelvű anyaga és eszköztára szabadon hozzáférhető, a korpusz bővíthető, tehát az alapszabályok betartása mellett bárki közzéteheti rajta keresztül a saját gyűjteményét. A CHILDES a TalkBank-rendszer részeként működik: https://childes.talkbank.org.

2. A CHILDES nemzetközi adatbázis és eszköztár

A CHILDES (Child Language Data Exchange System, Bernstein Ratner – MacWhinney 2016, 2018; Sagae et al. 2010) projekt 1984-ben indult Brian MacWhinney pszichológiaprofesszor szervezésében a Carnegie Mellon egyetemen. MacWhinney a doktori disszertációját óvodáskorú gyerekek morfológiai fejlődéséből írta, és ehhez gyűjtött spontánbeszéd-adatokat. A magyar nyelv gazdag morfológiája Magyarországra vonzotta, ahol öt óvodás magyar gyerekkel (Andi,

Éva, Gyuri, Móni, Zoli) készített felvételeket. Hazatérve az Egyesült Államokba, MacWhinney más kutatók gyűjteményeit is összeszedte, és a gyermeknyelvet kutató kollégái, valamint egy programozócsapat közreműködésével kidolgozott egy egységes átírási rendszert, és kifejlesztett számítógépes elemzőeszközöket a Unix operációs rendszerre. Mára a CHILDES adatbázisa 130+ különböző gyermeknyelvi korpuszt tartalmaz 30+ különböző nyelvből online, Windows-, Mac OS- és Unix-alapú elemzőprogramokkal együtt.

2.1. Az adatbázis

A CHILDES adatbázisában 2–9 éves, tipikusan fejlődő gyermekek spontán beszédén van a hangsúly, bár ennél fiatalabb és idősebb gyermekektől származó adatok is előfordulnak. Spontán beszéd alatt itt informális, kötetlen, a gyermek megszokott környezetében és megszokott családtagjaival zajló beszélgetést kell érteni. Ebben a kategóriában a korpuszok nagy része longitudinális, tehát nyomon követhetjük egy-egy gyermek nyelvi fejlődését éveken át, havi vagy akár heti rendszerességgel készült felvételeken keresztül. Az eredeti MacWhinney-féle és a később, mások által gyűjtött magyar adatok is ebbe a kategóriába tartoznak. A gyermekek többsége egynyelvű, de bilingvális korpuszok is találhatók az adatbázisban. A második legnagyobb kategória, ami a korpuszok összméretében messze elmarad a spontán beszédtől, a történetmesélés, azon belül is a Békamese elnevezésű, képekből álló történet elmeséléséről készült felvételek átirata. A Békamese- (angolul Frog Story) korpuszok Mercer Mayer amerikai gyerekkönyv író és illusztrátor 1969-ben megjelent munkáját, a Frog, where are you? (Béka, hol vagy?) című, 29, szavak nélküli, fekete-fehér vonalrajzból álló kalandtörténetét használják, amit a gyerekeknek a saját szavaikkal kell elmesélniük. A Békamese-korpuszoknak régi hagyománya van az angolszász gyermeknyelvkutatásban, ami már sok más régióra is kiterjedt. Jelenleg 13 nyelven találunk Békamese-korpuszokat a CHILDES adatbázisában. A morfológiai és szintaktikai fejlődés mellett a diskurzusjegyek és szövegkoherencia használatának vizsgálatára különösen alkalmasak ezek az adatok. A korpuszok harmadik kategóriája klinikai populációkkal készült felvételekből áll. Itt többek között Down-szindrómával, autizmussal vagy specifikus nyelvi zavarral élő, halláskárosult vagy agysérült gyerekekkel készült felvételek találhatók.

Az adatbázis a fenti kategorizáción túl nyelvenként (vagy nyelvcsaládonként), egy-egy nyelven belül pedig az adatgyűjtő kutató neve szerint rendeződik:

https://childes.talkbank.org/browser. A magyar adatokat az „Other” nyelvkategóriában találjuk a baszk, az észt és más magányos nyelvek társaságában. Három magyar korpusz található itt: az eredeti MacWhinney-korpusznak egy kibővített változata (MacWhinney 1974), Réger Zita longitudinális gyűjtése egy gyermekkel (Réger 1986), amely anyag a Nyelvtudományi Intézet jóvoltából még bővül, és Bodor Péter hasonló longitudinális korpusza egy kétéves gyermek első szavainak és mondatainak alakulásával (Bodor–Barcza 2007).

Az adatbázis elérhető online, de le is tölthető. A korpuszok három formában jelenhetnek meg az adatbázisban. A legegyszerűbb forma a hangfelvétel szöveges átirata – ez volt az eredeti, és sokáig egyetlen lehetséges mód, technikai korlátok miatt. A letölthető átirat a legegyszerűbb szövegfájlolvasóval is olvasható. Ma már lehetőség van az átirat és a hang összekapcsolására (a felhasználó hallja a hangot, és azzal párhuzamosan, a hangot valós időben követve látja az átírt szöveget), sőt az átirat, a hang és a mozgókép összekapcsolására is. Az adatbázis jelzi, hogy egyes korpuszok esetében elérhető-e hang-, illetve videófelvétel. A hanggal vagy videóval összekapcsolt átiratokat az online adatbázisban lehet követni, vagy letöltve a CHILDES szerkesztőszoftverével, a CLAN programmal lehet hallgatni, illetve szerkeszteni. A CLAN Windows, Mac és Unix operációs rendszerre is elérhető: http://dali.talkbank.org/clan. A későbbiekben visszatérünk még a használatára.

Bármely kutató kérheti a korpusza felvételét az adatbázisba, feltéve, hogy eleget tesz a CHILDES formai és etikai előírásainak. Ezek közül az egyik legfontosabb, hogy az adatgyűjtéshez etikai engedély szükséges, aminek a beszerzése történhet a kutatásvezető egyetemének vagy kutatóhelyének etikai szabályai szerint. A másik legfontosabb feltétel az, hogy a felvételeken szereplő gyerekek anonimizálva legyenek. Ez azt jelenti, hogy a felvételekből ki kell vágni minden olyan megnyilvánulást (vezetéknevet, címet, óvoda nevét stb.), ami a gyermeket azonosíthatja. Az etikai feltételek egyik következménye, hogy a videófelvételeket ritkán publikálják kutatók a nyilvános adatbázisban. A CHILDES természetesen a felhasználókat is felhívja a kutatóközösségben megszokott etikai normák betartására. A korpuszok szabadon hozzáférhetőek, de illik hivatkozni magára a CHILDES-ra és a felhasznált korpusz készítőjének megadott publikációira.

2.2. A CLAN program és az átírás szintaxisa

A hangfelvételek átirata és annotációi a CHAT formátumot követik. (Sajnos többszöri próbálkozással sem sikerült kiderítenem, hogy minek az akronimája vagy rövidítése lehet a CHAT betűsor.) Az átiratok .cha kiterjesztésű szövegfájlok a CHAT markereivel ellátva, amiket automatikusan konvertálni lehet más népszerű formátumokba, mint például Praat és ELAN. A szövegfájlokat bármilyen szövegszerkesztővel meg lehet nyitni, de az átíráshoz és annotációhoz érdemes a CHILDES saját szerkesztőprogramját, a CLAN szoftvert használni. A CLAN három fő üzemmódban működik: a chat mód az átírást könnyíti meg különböző funkciókkal, a sonic mód az átírás és hang/videó összehangolását, a coder mód pedig az annotációt. A CHAT és a CLAN részletes leírása megtalálható a neten: https://talkbank.org/manuals/CHAT.pdf.

A CHAT formátum a beszélt szövegek átírása mellett azok annotációjára is lehetőséget ad. A CHAT három típusú sort definiál: az egyik az átirat általános jellemzőit adja meg, és a `@` karakter vezeti be. A másik azt jelzi, hogy beszéd átiratát tartalmazza a sor. Ezt a `*` karakter és a beszélő hárombetűs kódja vezeti be. Egy beszélősorban csak egy mondat átirata szerepelhet, tehát minden mondat külön sorba kerül. Aki már próbált beszélt nyelvet átírni, nagyon jól tudja, hogy mennyire nem triviális kérdés a folyamatos beszéd mondatokra bontása. Praktikus okokból (az elemzés megvalósíthatósága miatt) azonban ez egy szükségszerű lépés. A harmadik sortípus a szöveg beszédsoronkénti annotációjára ad lehetőséget. Ezek a sorok egy-egy beszédsort követnek, és a `%` karakter és egy azonosító kód vezeti be őket. Az annotáció tartalmazhat morfológiai, szintaktikai vagy fonológiai elemzésen túl szociolingvisztikai markereket – vagy bármilyen egyéb megjegyzést, amit az átíró rögzíteni kíván. A sort bevezető azonosító kód jelzi, hogy milyen jellegű annotáció szerepel benne. Az (1) rövid részlet Réger Zita korpuszából jól illusztrálja a rendszert.

Az első két sor azonosítja a fájlt a CHILDES rendszerben. A `@Begin` kód jelzi, hogy kezdődik az átirat. A következő néhány `@` jelű sor megadja a nyelvet, a beszélők kódját és szerepét, a gyerek életkorát (2 év, 0 hónap, 25 nap), a hangfájl elérhetőségét, a dátumot, a hangfelvétel helyét a magnószalagon, az átíró nevét és a beszélgetés alaphelyzetét. Ezután következik a beszéd átirata, mondatonként. Ebben a részletben két annotációs sor van: a `%com` nevű megjegyzéseket tartalmaz arra vonatkozóan, hogy milyen helyzetbe került a gyerek, amikor éppen az adott

mondatot mondta, a %act nevű sor pedig a gyerek cselekedeteit írja le. Az átirat a @End kóddal végződik.

(1)

```
      @Loc: Other/Hungarian/Reger/020025.cha
1     @PID: 11312/c-00027754-1
2     @Begin
3     @Languages: hun
4     @Participants:    CHI Target_Child , MOM Mother
5     @ID:  hun|Reger|CHI|2;00.25||||Target_Child|||
6     @ID:  hun|Reger|MOM|||||Mother|||
7     @Media:     020025, audio, unlinked
8     @Date:      11-OCT-1992
9     @Tape Location:   Tape X , Side A. 212.
10    @Transcriber:     Szilvia Papp.
11    @Situation: Miki has just woken up. He is still in
bed.
13    *CHI: kivesz .
14    *CHI: anyu kivesz .
15    *CHI: anyu kivesz .
16    *CHI: anyu ki [//] anyu emej .
17    *MOM: jó kiveszlek , emellek .
18    *MOM: szervusz .
19    *MOM: jó reggelt !
20    *CHI: miki ajutt [=? aludt] .
21    %com: mother has lifted him out of the bed , miki is
sitting in her lap .
22    *MOM: hol aludt miki ?
23    *CHI: itt .
24    %act: points at bed .
...
31    *CHI: mag(n)ót .
32    *CHI: itt ?
33    %act: looking for the tape recorder .
...
903   @End
```

Angol (és néhány más) nyelvű szövegek morfológiai és szintaktikai elemzését automatikusan végzi a CHILDES. A morfológiai elemzés eredményét a %mor sorba, a szintaktikai elemzés eredményét pedig a %gra sorba illeszti. Magyar

nyelvre sajnos jelenleg egyik automatikus elemzés sem elérhető a CHILDES rendszerben, de lásd később a MONYEK korpusz leírását erre vonatkozóan. Egy angol példa a morfológiai és szintaktikai elemzésre a Manchester-korpuszból (Theakston et al. 2001):

(2)

```
@Situation: Structured Play
13    *CHI: I turned the cooker on .
14    %mor: pro:sub|I v|turn-PAST det:art|the n|cook&dv-AGT
prep|on .
15    %gra: 1|2|SUBJ 2|0|ROOT 3|4|DET 4|2|OBJ 5|2|JCT
6|2|PUNCT
16    *MOT: well done .
17    %mor: co|well part|do&PASTP .
18    %gra: 1|2|COM 2|0|ROOT 3|2|PUNCT
19    *MOT: cooking my pretzel ?
20    %mor: n:gerund|cook-PRESP det:poss|my n|pretzel ?
21    %gra: 1|0|INCROOT 2|3|MOD 3|1|OBJ 4|1|PUNCT
22    *CHI: it's cooked it already .
23    %mor: pro:per|it~aux|be&3S part|cook-PASTP pro:per|it
adv|already .
24    %gra: 1|3|SUBJ 2|3|AUX 3|0|ROOT 4|3|OBJ 5|3|JCT
6|3|PUNCT
```

A CHAT formátum szabályait és a megengedett annotációs (%) sorokat a depfile.cut nevű fájl tartalmazza, ami a CLAN programmal együtt telepítődik a számítógépre, ha a CLAN helyi, telepített változatát használjuk. Ez egy szövegfájl, ami a CLAN programmal (vagy bármely más szövegfájl szerkesztővel) szerkeszthető. Ez az, ami igazán rugalmassá teszi a rendszert, hiszen mindenki a maga igényei szerint definiálhatja az átirat szabályait. Az annotációs sorok kódjait is tetszés szerint definiálhatja a felhasználó további .cut fájlokban, ami lehetővé teszi például a magyar morfológiai elemzés beillesztését. Ezek a módosított, saját egyéni igények szerint kialakított kódrendszerek és .cut fájlok természetesen nem kerülnek be az online adatbázisba.

2.3. Az elemzőprogramok

A CLAN szerkesztő és az online adatbázis egy sor elemzőprogramot is kínál. A programokat parancssorral lehet behívni a CLAN szerkesztő egyik ablakában

vagy az online adatbázis bármelyik szintjén (a bal alsó sarokban). Az offline és online változat használata tökéletesen megegyezik egymással azzal az egy különbséggel, hogy míg az offline parancssorban specifikálni kell a mappát, amiben az elemzendő célfájlok vannak, addig az online parancssor automatikusan abban a mappában keresi a fájlokat, ahol éppen jár a felhasználó. A parancsok nem túl bonyolultak, a Unix operációs rendszer logikáját követik. A parancssor az elemzőalgoritmus nevével kezdődik, amit egy sor opció követ, majd az elemzendő fájlok neve zárja a parancsot. Ha azt szeretnénk megtudni például, hogy milyen b betűvel kezdődő szavakat milyen gyakorisággal használ a gyermek, a (3) parancsot adhatjuk:

(3) `freq -t% +t*CHI +s"b*" +u +o *.cha`

Itt a `freq` annak a parancsnak a neve, ami gyakorisági információt ad. A két `t` opció azt mondja, hogy ne az annotációs sorokban keresse a parancs a szavakat, hanem a gyermek beszédsorában (a sortípusokat tier-nek hívja a CHILDES, innen jön a `t`). Az `s` a keresendő string-et specifikálja: ebben az esetben a `b` betűvel kezdődő szavakat (egymástól nem-szó karakterrel elválasztott egységeket). Az `u` opció összegzi a fájlok keresésének eredményeit, az `o` opció gyakoriság szerinti sorrendben írja ki az eredményeket, a `*.cha` pedig a mappában található valamennyi .cha kiterjesztésű fájlt nevezi meg a keresés céljául. A parancs kimenetének első néhány sora MacWhinney Zoli-korpuszából a (4) alatt látható:

(4)

```
freq -t% +t*CHI +s"b*" +u +o *.cha
Thu Jan 24 12:09:19 2019
freq (27-Apr-2018) is conducting analyses on:
  ONLY speaker main tiers matching: *CHI;
****************************************
Speaker: *CHI:
474 bácsi
 46 bogár
 26 be
 23 bács
 18 bá
 11 bácsinak
 11 ba
```

```
 11 bemegyek
 11 bumm
  9 bírom
  9 baboda
```

Egy másik hasznos parancs a `kwal` (keyword and line), ami egy adott kulcsszónak a kontextusát mutatja meg. Ha például a fenti eredmény alapján azt szeretnénk megtudni, hogy miért mondogatta annyit a kétéves gyermek a *bírom* szót, az (5) szerint tehetjük meg:

(5) `kwal -t% +t*CHI +s"bírom" -w2 +w2 *.cha`

ahol a `-w` és `+w` azt specifikálja, hogy a célstring előtt és után két sort írjon ki a program. Az eredménynek egy részlete:

(6)

```
kwal -t% +t*CHI +s"bírom" -w2 +w2 +u *.cha
Thu Jan 24 12:25:41 2019
kwal (27-Apr-2018) is conducting analyses on:
  ONLY speaker main tiers matching: *CHI;
****************************************

From file "011000.cha"
----------------------------------------
*** File "011000.cha": line 401. Keywords: bírom, bírom
*CHI: jött (.) motor ott [% magának susog] ott teszem (.)
teszem (.) bujj el a róka kóma (.) itt van (.) bu bu (.)
itt is van (.) ott is van (.) kutyus .
*CHI: &=whisper (..) Barna bácsi , homokot [: homokba]
ülünk .
*CHI: együnk [//] egyetünk (.) itt itt [!] a nagy autó
(.) ott megyünk (.) alig bírom másik [% phrase] (.) &o
itt van a Vio [!] néni (.) autóval megyünk (.) alig
bírom .
*BRI: Zoli , tudod miért nem megy mert ki van szedve
belôle az elem (.) azért nem megy az autó .
*CHI: Barna bácsi (.) Barna bácsi .
```

```
From file "011001.cha"
----------------------------------------
*** File "011001.cha": line 552. Keyword: bírom
*CHI: fütyöl [: füstöl] .
*MON: egér (.) itt (.) egér (.) csüccsülj le , egér .
*CHI: alig bírom .
*BRI: ajaj , alig bír .
*CHI: Barna bácsi , dolgozunk .

From file "011002.cha"
----------------------------------------
*** File "011002.cha": line 544. Keywords: bírom, bírom,
bírom, bírom
*CHI: várj csak , Barna bácsi (.) várj csak .
@New Episode
*CHI: nagy (.) indulás (.) (.) alig bírom (.) elszakadok
(.) alig bírom (.) alig bírom (.) elszakadok (.) alig
bírom (.) elszakadok .
*UNK: hadd nézzem ezt a halacskát .
*CHI: jó itt (.) elromlott a +...
```

2.4. A CHILDES egyéb eszközei: a LuCiD Toolkit

A LuCiD Toolkit egy brit gyermeknyelv kutató társulás (az ESRC International Centre for Language and Communicative Development) eszköztára, melyet a CHILDES adatbázisának hatékonyabb kihasználására fejlesztett ki (Chang 2017). Az eszközök elérhetők a CHILDES oldalairól: http://gandalf.talkbank.org:8080. A CHILDES Browser lehetővé teszi a teljes adatbázis célirányos keresését bizonyos kritériumok szerint, mint például korpuszméret, a mondatok hossza vagy a gyermek kora. A Restricted Distribution eszköz feltérképezi azokat a szavakat, amelyek egy adott, reguláris kifejezéssel specifikált kontextusban előfordulnak. Megmutatja például, hogy a Bodor-korpuszban a mondatok legnagyobb valószínűséggel az *és, hát, nem, na, mit* vagy *igen* szóval kezdődnek. A CHILDES Generator megbecsüli egy adott kifejezést követő szavak bigram valószínűségét a korpusz alapján, a Distributional Word Classification eszköz pedig ngram valószínűségeket von ki a korpuszból és disztribúciós tulajdonságaik szerint rendezi a célszavakat.

3. A magyar gyermeknyelvi korpuszok

3.1. A MONYEK

A MONYEK (Magyar Óvodai Nyelvi Korpusz), Mátyus Kinga doktori disszertációjához készült hangfelvételekből és azok átiratából áll (Mátyus–Orosz 2014; Orosz–Mátyus 2014). A korpusz 62 óvodás gyerekkel folytatott interjú alapján készült; összesen mintegy 140 000 szót tesz ki. A gyerekek kiválasztásánál fontos szempont volt a társadalmi-gazdasági státuszuk. A gyerekek egyik felét a KSH adatai szerint elit budapesti óvodákból toborozták a kutatók, a másik felét pedig alacsonyabb társadalmi-gazdasági státuszú budapesti óvodákból. Az interjúk előre meghatározott forgatókönyv szerint zajlottak, az erősebben kontrollált feladatoktól a szabadabb társalgási feladatok felé haladva. Az első feladatban a kísérletvezető – képek alapján – elmondott egy történetet (Zsuzsi és az állatok), amit a gyermeknek vissza kellett mondania. Ezt követte egy önálló képmesélési feladat, ahol a fent már említett Békamesét és két rövidebb történetet meséltek el a gyerekek képek alapján. Ezután egy játék (pl. foci) menetét és szabályait írták le a gyerekek, és végül irányított beszélgetés zárta az interjút.

A hanganyag átírása a CHILDES CHAT formátumának szabályai szerint történt; a szöveg automatikus morfológiai elemzését a HuMor végezte, melynek kimenetét a PurePos egyértelműsítő rendszer gyereknyelvre adaptált változata egyértelműsítette, utólagos kézi ellenőrzéssel. A korpusz elérhető CHAT fájlok formájában és xml-formátumban. A MONYEK-korpusz készítését a CESAR-projekt és az MTA Nyelvtudományi Intézet támogatta.

3.2. A GABI

A GABI (Gyermeknyelvi beszédAdatBázis és Információtár) Bóna Judit irányításával készül az ELTE Fonetikai Tanszékén (Bóna 2017, Vakula–Váradi 2017). A korpusz különlegessége, hogy minden korosztályból gyűjt hangadatokat a 3 évesektől egészen a 18 évesekig, és így szinte minden nyelvfejlődési korszakot lefed. A korpusz értelemszerűen keresztmetszeti, tehát nem ugyanazokat a gyerekeket követi végig, hanem különböző gyerekektől gyűjt adatokat a különböző korosztályokban. A gyermekek kiválasztásának szempontja, hogy köznyelvet beszéljenek, és neurotipikus fejlődésűek legyenek. Részletes anamnézist vesznek fel velük, amelyben rákérdeznek a családi háttérre (szülők iskolázottsága, változás a családszerkezetben, anyagi helyzet, testvérek stb.), óvodára, iskolára, betegsé-

gekre, allergiára, nyelvfejlődési jellemzőkre (pl. járt-e a gyermek logopédushoz, mikor kezdett beszélni stb.).

Az adatfelvétel forgatókönyv szerint zajlik a BEA (BEszélt nyelvi Adatbázis, MTA Nyelvtudományi Intézet) elnevezésű felnőtt beszélt nyelvi korpusz módszertanát követve, és azt gyerekekre adaptálva. A feladatok egy része erősen kontrollált (mondatismétlés, mondatolvasás), más része valamivel több szabadságot ad (szavak definiálása, hallott történet tartalmának visszamondása), és végül kvázi-spontán beszédet kiváltó feladatokat is találunk (történetmondás képek alapján, szabad beszélgetés egy adott témáról). A forgatókönyv részletei a gyermek korának megfelelően kisebb mértékben változnak.

E tanulmány megírásáig több mint 450 gyermekkel készült felvétel, melyek hanganyagának átírása és annotálása folyamatban van. Mint azt említettem, a GABI jelenleg nem hozzáférhető, de érdemes a kutatócsoport honlapját figyelni későbbi fejleményekért: http://fonetikaitanszek.elte.hu/index.php/kutatas/gabi.

Irodalom

Bernstein Ratner, N., MacWhinney, B. 2016. Your laptop to the rescue: Using the Child Language Data Exchange System archive and CLAN utilities to improve child language sample analysis. *Seminars in Speech and Language* 37. 74–84. https://psyling.talkbank.org/years/2016/ssl/nan.pdf

Bernstein Ratner, N., MacWhinney, B. 2018. TalkBank resources for psycholinguistic analysis and clinical practice. In: Pareja-Lora, A., Blume, M., Lust, B. (szerk.). *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*. Cambridge, MA: MIT Press. https://psyling.talkbank.org/years/2019/RatnerMacW.pdf

Bodor P., Barcza V. 2007. Acquisitioin of diminutives in Hungarian. In: *The acquisition of diminutives: A cross-linguistic perspective*. Amsterdam: Benjamins. 231–263.

Bóna J. 2017. GABI – Gyermeknyelvi beszédadatbázis a kutatásban. In: Bóna J. (szerk.). *Új utak a gyermeknyelvi kutatásokban*. Budapest: ELTE Eötvös Kiadó. 35–50.

Chang, F. 2017. *The LuCiD language researcher's toolkit* [Computer software]. http://gandalf.talkbank.org:8080/

MacWhinney, B. 1974. *How Hungarian children learn to speak*. Unpublished doctoral dissertation. Berkeley: University of California.

Mátyus K., Orosz Gy. 2014. MONYEK: Morfológiailag egyértelműsített óvodai nyelvi korpusz. *Beszédkutatás* 22. 237–245.

Orosz Gy., Mátyus K. 2014. An MLU estimation method for Hungarian transcripts. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. *Text, Speech, and Dialogue, of Lecture Notes in Computer Science*. Vol. 8655. 173–180.

Réger Z. 1986. The functions of imitation in child language. *Applied Psycholinguistics* 7/4. 323–352.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., Wintner, S. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language 37*/3. 705–729. DOI: 10.1017/S0305000909990407. https://psyling.talkbank.org/years/2010/jcl-sagae.pdf

Theakston, A. L., Lieven, E. V. M., Pine, J. M., Rowland, C. F. 2001. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language* 28. 127–152.

Vakula T., Váradi V. 2017. Gyermeknyelvi hangfelvételek rögzítésének és lejegyzésének tapasztalatai. In: Bóna J. (szerk.). *Új utak a gyermeknyelvi kutatásokban*. Budapest: ELTE Eötvös Kiadó. 51–64.